Workshop opérations 2011
Journée Cloud
21/10/2011

Présents

- Fabienne Anhalt (Lyatiss)
- Jean-Michel Barbet (SUBATECH)
- Cécile Barbier (LAPP)
- Catherine Biscarat (LPSC)
- Christophe Blanchet (IBCP)
- Chmouel Boudjnah (Rackspace)
- Carlos Carranza (CPPM)
- Tony Cass (CERN)
- Frédérique Chollet (LAPP)
- Nicolas Clémentin (LUPM)
- Alexandru Costan (INRIA)
- Jean-Claude Chevaleyre (LPC)
- Hélène Cordier (CCIN2P3)
- Stéphane Delmotte (LBBE/PRABI)
- Frédéric Desprez (INRIA)
- Michèle Detournay (APC)
- Christophe Diarra (IPNO)
- Christelle Eloto (CCIN2P3)
- Joseph Emeras (INRIA)
- Eric Fede (LAPP)
- Jacques Garnier (CCIN2P3)
- Clément Gauthey (IBCP)
- Pierre Gay (Université Bordeaux)
- Pierre Girard (CCIN2P3)
- Christine Gondrand (LPSC)
- Haïkel Guémard (Sysfera)
- Michel Jouvin (LAL)
- Edith Knoops (CPPM)
- Adrien Lèbre (Ecole des Mines)
- Christine Leroy (IRFU)
- Charles Loomis (LAL)
- Liliana Martin (LPNHE)
- Gilles Mathieu (CCIN2P3)
- Emmanuel Medernach (LPC)
- Jean-Pierre Meyer (CEA/IRFU)
- Geneviève Moguilny (IPGP)
- Jérôme Pansanel (IPHC)
- Guillaume Philippon (LAL)
- Mattieu Puel (CCIN2P3)
- Denis Pugnère (IPNL)
- Pierre Riteau (INRIA/IRISA)
- Geneviève Romier (IDGC)
- Albert Shih (OBSPM)
- Floris Sluiter (SARA)
- Frédéric Suter (CCIN2P3)
- David Weissenbach (IPGP)

Agenda : http://indico.in2p3.fr/conferenceTimeTable.py?confId=5551#20111021.detailed

Introduction

Adrien Lèbre, Ecole des Mines de Nantes
http://indico.in2p3.fr/contributionDisplay.py?contribId=10&confId=5551

Adrien lèbre a lancé en 2008 un groupe de travail dans Grid5000 autour de la virtualisation.

AL présente le cloud comme la rencontre du monde *internet* avec le monde *computing* puis présente les différentes déclinaisons cloud (IaaS, PaaS, SaaS), notant que le SaaS fait partie du paysage grand public depuis plusieurs années déjà (exemple donné de l'entrerise Porsche dont la totalité du Système d'Information est fournie par Google).

Quelques points de la présentation :

- Démonstration des limites et contraintes de l'utilisation grille en mode classique, en comparaison avec une gestion virtualisée (« The Alice and Bob example »)
- Principe de virtualisation et caractéristiques de la gestion de VM
- Principes de l'utilisation IaaS et panorama des solutions OpenSource disponibles
- Principe de fédération de clouds (« sky computing »)

La conclusion d'Adrien est que les données sont probablement le point critique, vis-à-vis de la performance, la sécurité et la fiabilité.

Questions/discussion

F.Chollet : Est-on prêts techniquement à garantir un accès 24/7 pour les utilisateurs ?
AL : Le stockage dans le cloud n'est pas encore tout à fait fiable, sauf à mettre en place du stockage persistent qui induit un coût supplémentaire pour garantir une certaine fiabilité

N.Clémentin : pour estimation, a-t-on une idée du manpower nécessaire par VM ?
AL : il n'y a pas de chiffre défini. Ce qu'il faut réaliser, c'est que le travail ne décroit pas, puisqu'au final le nombre de machines physiques reste constant. C'est la capacité qui augmente avec une même infrastructure.

G.Mathieu : Les pertes de performances liées à la virtualisation sont-elles chiffrées ?
AL : pour l'I/O on note une perte moyenne de 20%. Mais c'est à mettre en regard des gains sur d'autres aspects, par exemple pour Mapreduce et HPC, avoir un accès à de nombreuses ressources CPU équilibre les pertes en I/O.
Ce chiffre varie aussi selon l'hyperviseur utilisé, en sachant que pour certains (VMWare par exemple) les chiffres de performance ne sont pas communiqués.

Building an Openstack cloud

Chmouel Boudjnah, Rackspace
http://indico.in2p3.fr/contributionDisplay.py?contribId=14&confId=5551

Chmouel Boudjnah travail sur Openstack qu'il présente comme une pile de différents produits.

CB commence par un historique d'Openstack, débutant en 2010 par un projet cloud porté par Rackspace ensuite tourné vers l'*Open Source*.

Quelques points de la présentation :

- Notion de *object storage* : ce n'est pas un système de fichiers (structure de type Account/container/object). Similaire à la solution Amazon S3. Basé sur une API REST avec fonctions de base (get, put, delete…). Système construit pour être le plus efficace possible en limitant les coûts : basé sur du matériel relativement bon marché, avec toute la redondance faite au niveau du software.
- Evolution du *object storage* d'une architecture avec base de données centralisée vers une infrastructure distribuée (passage de "rackspace cloud files" en 2008 à "Openstack open storage" en 2010)
- Principes et avantages de la gestion du stockage physique via un *objects ring* (principe de *consistent hashing* – mapping nom/entités)

Questions/discussion

P.Girard : Y a-t-il des limites dans la taille des fichiers ?
CB : Limite de 5Go. Pour les objets plus larges, le découpage est à faire par l'utilisateur mais le système se charge de re-concaténer l'info à la lecture. Il existe des outils qui se chargent de faire ce découpage. Les accès se font de manière séquentielle

P.Girard : peut-on faire du remote I/O ou est-on obligé de télécharger l'ensemble des fichiers ?
CB : Il existe un système permettant de faire du remote I/O.

A.Lèbre : Gérez-vous le principe de fédération dans le design du Swift ?
CB : le cloud bursting et la fédération sont des problématiques importantes pour OpenStack. Il y a dans swift un module de synchronisation qui permet de synchroniser les données distribuées au niveau du conteneur. HP commence à implémenter un principe de distributed datacenter, nous ne l'avons pas encore mais c'est une solution qui se développe.

Data management in clouds: the BlobSeer approach

Alexandru Costan, INRIA
http://indico.in2p3.fr/contributionDisplay.py?contribId=15&confId=5551

Alexandru Costan est PhD researcher at INRIA et travaille sur le projet BlobSeer.

AC commence par expliquer les principes de gestion de données dans le cloud, et présenter les solutions commerciales disponibles. Le point fort de ces solutions est la haute disponibilité, le point faible est souvent la limitation en taille des fichiers stockés.

Quelques points de la présentation :

- Présentation de BlobSeer : plateforme de gestion d'objet de type BLOB (*Binary Large Objects*). Description des choix techniques et de l'architecture
- Principe d'utilisation du *versioning* pour gérer la concurrence d'accès aux données
- Notion de *checkpoint-restart* (par opposition à réplication) et des principes de *cloning* et *shadowing.*
- Présentation de cas d'utilisation
- Intégration dans Nimbus en projet

Questions/discussion

M.Puel : Envisagez-vous d'utiliser BlobSeer dans d'autres projets ?
AC : On va essayer avec OpenNebula, notamment via une participation commune à un projet européen (Scalus). Dans Nimbus nous avons remplacé le système de stockage « cumulus » par BlobSeer. Nous n'avons pas encore essayé avec OpenStack mais nous y arriverons sans doute dans le futur.

A.Lèbre : Comment BlobSeer peut prendre en compte les principes de burst, notamment pour passer d'un cloud à un système de donnée ?
AC : Pour accéder aux données ce sera par interface.
A.Lèbre : Donc il faut actuellement copier les données dans BlobSeer ? Comment gérer la synchronisation ?

AC : On l'a fait pour certains produits (remplacement de HDFS pour adobe) mais pour l'instant il n'y a pas d'outil permettant de faire cette interface avec d'autres systèmes de fichiers.

Trusted VM Images: the HEPiX Point of View

Tony Cass, HEPiX virtualization working group, CERN
http://indico.in2p3.fr/contributionDisplay.py?contribId=11&sessionId=1&confId=5551

Tony travaille au CERN et coordonne le *HEPiX virtualisation working group*

TC commence par répondre à la question de départ : où est le problème ? Pourquoi ne peut-on pas faire comme Amazon qui fait confiance aux images lancées ? Les réponses avancées sont la paranoïa, le manque d'isolation du hardware sur lequel tournent les VMs, et le besoin de traçabilité.

Quelques points de la présentation :

- Présentation de la solution HEPiX de *Trusted Virtual Images*
- Politiques pour la génération de *trusted images*
- *Image contextualisation* et support multi-hyperviseur
- Liens avec le marketplace StratusLab

Question ouverte : est-ce que cet aspect de la virtualisation nous rapproche du cloud ?
Possibilité pour les sites d'instancier des VMs qui se connectent directement aux infrastructures de soumission des VOs. C'est un pas vers le cloud pour la communauté HEP

Questions/discussion

P.Girard : la plupart des sites ont des batch systems, n'a-t-on pas une transition où on devrait faire tourner nos clusters en utilisant le batch system pour faire tourner des images ?
TC : ce serait logique de faire de cette façon. On peut imaginer que plutôt d'avoir une queue de 140000 jobs on ne gère que 80000 machines virtuelles. Je ne suis plus impliqué dans la façon dont on va le faire
E.Fede : le batch local est aussi utilisé pour jongler entre les engagements sur les ressources
TC : la solution de virtualisation est plus simple car elle permet de gérer les engagements en terme de capacité

J**PMeyer** : que se passe-t-il si un site n'est pas pur HEP, et que se passe-t-il si une VO n'utilise pas ses ressources ?
TC : si les ressources ne sont pas utilisées on est plus dynamiques pour allouer les ressources disponibles en mode virtualisation. Il y aura toujours des problèmes si les ressources ne sont pas utilisées sur une longue période et si l'utilisateur veut les utiliser intensivement par la suite pendant une longue période également.
Si tout tourne à plein régime et que je dois maintenir un équilibre, je laisse le système avec les règles définies. Si une partie des utilisateurs n'est pas active on peut donner plus aux autres.

Grid5000 – constructing Software Environments

Joseph Emeras, INRIA – MESCAL team
http://indico.in2p3.fr/contributionDisplay.py?contribId=12&confId=5551

J.E. commence par revenir à la question initiale : Pourquoi utiliser des images d'environnement ?

Quelques points de la présentation :

- Outils de déploiement d'environnements : Présentation de CFEngine, Puppet, Chef, Juju, UShareSoft, Kameleon. Revue de leur fonctionnement et des concepts associés.
- Présentation générale de Grid'5000 et de la gestion d'environnement dans le projet

Conclusion : nous sommes poussés à utiliser de plus en plus de VMs mais l'environnement utilisateur devient un peu une boite noire.

Questions/discussion

A.Lèbre : aujourd'hui, peut-on déployer physiquement un environnement sur grid5000 ?
J.E : Oui, il y a un mécanisme qui permet par exemple d'envoyer un environnement physique sur la machine qui va pouvoir instancier des VMs.

A.Lèbre : Une archive G5k peut-elle être instanciée en VM ?
J.E : C'est possible. Avec Kameleon, on peut faire de l'archive à la fois une VM et à la fois une image G5K

StratusLab Marketplace

Charles Loomis, LAL-IN2P3
http://indico.in2p3.fr/contributionDisplay.py?contribId=13&confId=5551

Cal Loomis est coordinateur du projet européen StratusLab.

CL commence par présenter de manière générale le projet Stratuslab, et son architecture IaaS, avant de se concentrer sur le système de gestion des images : le market place.

Quelques points de la présentation :

- Définition des principes et de la raison d'être du Marketplace
- Description des métadonnées, standards et outils
- Présentation des workflows, notamment la gestion des politiques de sécurité par les sites (possibilité pour les sites de définir les politiques d'autorisation pour une image donnée) et le partage d'images.

Questions/discussion

P.Girard : y aura-t-il un outil de sécurité dans stratuslab ? E.g. si on arrive à vous attaquer et à déposer des images qui ne sont pas saines, comment cela va-t-il être géré ?
CL : L'audit sécurité de l'infrastructure est prévue dans l'année 2 du projet. Par contre, il faut noter que Stratuslab n'est pas un organsime qui va faire les audits des images. On propose les outils mais on ne le fera pas car il n'y a pas de pérennité pour cela.

S.Delmotte : Si je dépose une image dans le marketplace, est-ce que j'aurai un retour sur cette image de la part de ceux qui vont l'utiliser, pour me pousser à la faire évoluer ?
G: l'idée est de permettre d'ajouter par ex. des commentaires et autres métadonnées sur les images. Si c'est un requirement important nous pouvons développer cela

T.Cass : est-ce que vous avez avancé pour la planification sur le support et la maintenance de logiciel après la fin du projet (2012)

CL : Cela commence a être discuté. Il y aura une discussion avec les différents groupes de production middleware (EMI, EDGI etc.) pour envisager la suite du support des produits lorsque ces projets arriveront à terme.

Migration de machines virtuelles

Pierre Riteau, INRIA/IRISA, Rennes
http://indico.in2p3.fr/contributionDisplay.py?contribId=16&confId=5551

PR commence par un point général sur la virtualisation du matériel, les concepts de *full virtualisation* et *paravirtualisation* et les principes associés, avant de se concentrer sur la migration.

Quelques points de la présentation :

- Idée de la migration : transfert de VM (et de son état) d'un hôte physique vers un autre
- Principes des approches *Stop-and-copy* et *live migration* et cas d'utilisation
- Techniques et algorithmes de *live migration* notamment *pre-copy* et *post-copy*
- La migration live sur un large réseau (WAN) et les possibilités d'optimisation (compression, *delta transfer*, dé-duplication de données)

Conclusion : la migration à chaud est toujours un sujet de recherche actif et il y a une volonté au niveau industriel de profiter des innovations liées. Il va être important de suivre Les travaux sur les couches au dessus de la migration.

Questions/discussion

S.Delmotte: Y a-t-il des outils permettant de migrer une machine physique vers une VM ?
PR : Chez Redhat il y a beaucoup d'efforts faits dans cette direction pour offrir des outils de gestion, dont ceux de migration d'une machine physique vers une machine virtuelle

G.Mathieu : Quelle est la proportion du temps nécessaire pour migrer une page mémoire par rapport à migrer un état CPU par exemple ?
PR : La migration CPU est vraiment minime puisqu'il s'agit simplement des registres, même si ça dépend beaucoup du type de matériel qu'on a. Au final c'est de l'ordre du Ko, donc vraiment négligeable par rapport à la mémoire.

HPC cloud – calligo system

Floris Sluiter, SARA, Amsterdam
http://indico.in2p3.fr/contributionDisplay.py?contribId=17&confId=5551

Floris Sluiter est consultant à SARA aux Pays-Bas et travaille sur différents projets, notamment BiG Grid et son projet de cloud HPC.

FS présente rapidement SARA, le contexte du projet de HPC cloud et quelles sont les motivations derrière un tel projet.

Quelques points de la présentation :

- Le départ du projet, lancé sous la forme d'une compétition d'idées.
- Présentation des utilisateurs du HPC cloud, des types d'applications

- Description de l'architecture de l'infrastructure *calligo*
- Analyse rapide de rentabilité : est-ce que le cloud est cher ?
- Aspect sécurité

Conclusion : facile à fournir du support, facile à utiliser. L'idée de fond est de changer l'environnement, pas les applications.

Questions/discussion

P.Girard: What is the easiest between administrating a grid site and administrating an HPC cluster?
FS: m/w part is quite complicated but for a skilled admin that is not a problem. User support is not a hard job.

M.jouvin: is the configuration done manually?
FS: This is all scripted. User creation scripts that create V-LAN, port forwarding, firewall rules etc. Some of them are dynamically started

P.Riteau: what is your experience with SIARB?
FS: Depending on the h/w you use not all servers are up to speed. We are now fine
P.Riteau: And what about latency differences between physical/cloud?
FS: It should go down to 10%, there is indeed an overhead but not too much

F.Chollet: Do you provide virtual machine images, or do the users come with their own?
FS: We have some templates, with endorsers etc.

P.Girard: User support is simpler, but at the starting point did you have to help your users get up to speed with virtualized infrastructure?
FS: People are used to software installation. Most of them are administrators. We do offer some templates so this help, but it usually gets very quickly.

A.Lèbre: You currently have 19 nodes/600 cores. What is your estimation of the size of the virtualized infrastructure in terms of nb of VMs?
FS: We are about 100 VMs now, where users can use 32 to 64 cores at a time. But we have no limitation really. All in one this is a small HPC cloud (amazon has around 1M cores). We currently have 30 active users and 25 waiting for the new hardware. Ideally we could expend up to 3k cores.

Scheduling HPC applications in the cloud

Yacine Kessaci, Université de Lille/CNRS/INRIA
http://indico.in2p3.fr/contributionDisplay.py?contribId=18&confId=5551

Yacine Kessaci est actuellement en thèse à l'université de Lille et travail sur les problématiques d'ordonnancement en vue d'économie d'énergie.

YK commence par présenter les motivations du travail en cours, notamment en le mettant en perspective par rapport à la volonté d'économie d'énergies.

Quelques points de la présentation :
- Architecture étudiée basée sur une fédération de clouds
- Problématique d'ordonnancement dans l'optique cloud, avec présentation des dimensions prises en compte (énergie, émissions de CO2, profit)

- Algorithmes et mécanismes utilisés
- Présentation des résultats et des conclusions du travail de recherche

Questions/discussion

JM.Barbet : Le prix de l'électricité change souvent au cours de la journée. Est-ce que vous prenez ça en compte dans le calcul et la fréquence des ordonnancements ?
YK : Il y a un ordonnancement toutes les 50s. Pour l'instant, on ne prend pas en compte les changements du prix mais ça peut être un axe d'amélioration en affinant le modèle.

A.Lèbre : Le scheduling est-il plutôt à la mode batch scheduling, ou bien avez-vous envisagé de pouvoir stopper un « job » pour libérer de la place aux autres ?
YK : On a travaillé plutôt sur le cerveau de l'ordonnanceur plutôt qu'un ordonnanceur particulier.
Tout ce qui a été décidé au moment de l'ordonnancement n'est pas modifié par la suite.

Expand Network virtualisation to the cloud

Fabien Anhalt, Lyatiss, Lyon
http://indico.in2p3.fr/contributionDisplay.py?contribId=19&confId=5551

Fabienne travaille à Lyatiss qui est une startup dont les activités sont centrées autour des problématiques de virtualisation du réseau.

FA présente tout d'abord le contexte général dans lequel se place la thématique abordée, et notamment quelques perspectives sur l'utilisation du cloud, et met l'accent sur la complexité sous-jacente du cloud.

Quelques points de la présentation :
- Présentation de la problématique derrière le concept de virtualisation réseau, et des risques que présente le fait d'ignorer l'aspect réseau dans le cloud.
- Concepts basiques de la virtualisation réseau : routeurs et liaisons virtuelles.
- Histoire de la virtualisation réseau
- Description du langage VXDL utilisé pour la configuration d'infrastructures de réseau virtuel
- Présentation des principes et du fonctionnement de CloudWeaver, solution logicielle proposée par Lyatiss.

Questions/discussion

P.Riteau : Openstack travaille sur une pile network as a service, comment Lyatiss se positionne ici ?
FA : On peut avoir un agent openstack pour partager et intégrer ces ressources, et avoir par ex. un router CISCO. L'idée est de permettre une infrastructure hétérogène qui puisse être gérée par CloudWeaver.
P.Riteau : Lyatiss n'est pas impliqué officiellement, n'avez-vous pas peur de concurrents qui prennent la place ?
FA : La concurrence est toujours possible.

?: Quelles sont les contraintes hardware pour Cloudweaver ?
FA : Un ensemble d'agents a été écrit pour les différents h/w. L'interface entre les agents et le h/w est assez simple.

E.Fede : au delà de l'agrégation des liens, peut-on faire avec un réseau virtuel tout ce qu'on peut faire avec un vrai réseau, comme IPv6 etc. ?
FA : On peut spécifier un contrôle clé, faire du routage à différents niveaux, on peut demander un certain équipement. Beaucoup de choses sont donc possibles

G.Romier : Est-ce qu'un équipement réseau peut-être utilisé à la fois en mode virtualisé et en mode physique ?
FA : La solution dans ce cas serait de faire deux profils virtuels, un pour l'administration et un pour l'utilisation

M.Puel : Au niveau des performances, est-ce que l'utilisation d'un réseau virtuel a un impact ?
FA : L'agent se met dans l'hyperviseur et pas dans la VM. Il n'a pas d'impacts sur les performances outre les impacts connus du overhead de la virtualisation déjà présenté. Avec Xen et KVM on peut avoir du 100% sur gros fichiers alors que si le mode de transfert est en rafales il y aura plus d'impacts. Donc ça dépend du nombre d'interfaces virtuelles à gérer. Avec l'augmentation des performances des machines, cela tend à diminuer

G.Romier : comment fonctionne le système de licence ?
FA : ne peux pas répondre exactement sur cet aspect.

Résumé de la journée

Ce qu'on a fait

- Démystifier le cloud. Après tout, c'est pas si obscur que ça…
- Partager beaucoup de connaissances, de concepts et d'exemples concrets.
- Faire ou consolider le lien entre les *virtualisation geeks* et les *grid neirds*. Même si ce sont quelquefois les même !

Ce qu'il faut faire maintenant

- Faire mûrir la réflexion, et laisser les idées faire leur chemin…
- Réfléchir à nos possibilités, ce que l'on pourrait/voudrait faire. Par ex : Le passage au cloud passe-t-il par une étape préliminaire de pure virtualisation ?
- Identifier les potentiels problèmes, limitations et challenges que représente la mise en place de solutions cloud dans notre contexte, vis-à-vis des thématiques abordées aujourd'hui (Confiance et sécurité, stockage, gestion des ressources, réseau)
- Commencer à cibler les utilisateurs potentiellement intéressés par une solution cloud
- Se préparer à aller au devant de ces utilisateurs